Tracking Users Like There is No Tomorrow: Privacy on the Current Internet

Iskander Sánchez-Rola, Xabier Ugarte-Pedrero, Igor Santos, and Pablo G. Bringas

S3lab, DeustoTech - Computing, University of Deusto, Bilbao, Spain
{iskander.sanchez,xabier.ugarte,isantos,pablo.garcia.bringas}@deusto.es

Abstract. Since the beginning of the web, users have been worried about usability but not always about security or privacy. Nowadays people are starting to realize that sometimes it is important to protect their privacy not only in real life, but also in the virtual world. This paper analyzes the current privacy debate surrounding online web tracking and explains the most relevant techniques and defenses. It also presents the different companies involved and related standards and regulations.

Keywords: web privacy, online tracking, information security

1 Introduction

Privacy is a right often violated in the current Internet, sometimes due to the ignorance of the users, and other times because of the abuse of service providers. Therefore, it has become an issue of great concern for users. According to the Oxford English Dictionary [1], privacy is the state of being free from public attention. Based on that premise and considering the technological environment in which we find ourselves, it is harder than ever to preserve that right [2, 3]. For this reason, privacy is more important than ever.

Online privacy goes far beyond accepting some terms and conditions in social networks [4, 5] but continually collecting a large amount of data with or without our permission [6]. The data collected is as diverse as browser identificators and browsing history [7]. This can happen for various reasons, such as not reading the privacy policies correctly, or simply because online advertisers collect more data than the strictly necessary. Although part of the information collected is not dangerous independently, if crossed, it can become a serious privacy invasion.

The intention of this paper is to familiarize computer security and privacy researchers with web tracking, examining and discussing all the different factors and privacy implications related to usual web browsing.

2 Privacy Attacks

Despite privacy-violating techniques are possible and even likely, sometimes we do not know how many different ways actually exist and to what extent are a threat. This section reviews the most common privacy attack vectors and explains the different techniques used.

2.1 Fingerprinting

Identifying someone unequivocally on the Internet is a common practice nowadays. Furthermore, fingerprinting allows to gather huge amounts of data related to user browsing, independently of where he is [8]. This technique raises serious privacy concerns for users.

All the data obtained could be used to protect users and web applications against malicious actors, for instance, by detecting the use of stolen credentials. However, it is also possible to use the information to conduct specific attacks against users. There are three main types of fingerprinting:

- **Browser recognition**: A fingerprint is a list of attributes that have disparate values between different web browsers but always have the same on each one. If those values are distinctive enough, their combination could be unique and work as an identifier [9]. These attributes consist of version number of the browser, screen resolution, and the list of used fonts, among others [10]. Canvas fingerprinting is another type of browser or device fingerprinting technique that leverages the Canvas API of the browser [11, 12], exploiting the differences in the rendering of the same piece of text in order to get an identifier.
- **Unique IDs**: Maybe the most effective and simple method of identification is creating a single javascript for each user, including a unique identifier in a variable. This javascript is cached and will always be used. Another interesting technique is to return images with a unique and exclusive Etags [13]. During the next connection, the server will realize that there has not been any change in the image, which implies that it is the same user. Identifiers can also be sent in HTTP requests using redirections or javascript for the assignation.
- **Cognitive identification**: JavaScript allows a website to easily create a full itinerary of all the interactions of the user with the different parts of the webpage just making use of event handlers of mouse and keyboard [14, 15]. Mouse moves, scroll-behavior and highlighted texts are some of the obtainable data that can be used to identify certain browsing patterns.

2.2 Information Storage

External code included in a website, has access to many different parts of the host website. This information, susceptible of being leaked, include cookies and many other sensitive data. Even if external code could enhance the user experience, it could also be used for malicious purposes.

- **Cookies**: These are the most common option (both Flash and HTTP). It is very easy to get information about the user's browsing habits with this method and to combine with user-identifying data [16]. For instance, a server can relate different identifiers from the users with the information in the referer header of the request, sharing cookie values between websites (i.e.,

syncing). Although cookies can be deleted, accepted or blocked, they are the cause of many online privacy attacks to the user. An example is to store an HTTP and a Flash cookie, and if the user removes the HTTP cookie, copy the value from the Flash cookie (i.e., respawning).
- **HTML5**: Using local storage, websites have the possibility of storing information in the browser of the user [17]. Before, the only way of storing data was with cookies. This method is more secure, and allows websites to store many information locally (more than 5MB), without slowing down browsing. Information is never transferred to the server and is domain dependent. All the sides from the same domain, can access the information or store new. This local storage technique presents the same problem as cookies.
- **Javascript**: Window.name, is a non-persistent property of Javascript (could be stored in cache), which is used to pass information between different website pages. Even if it is often used for setting targets for hyperlinks and forms, it has security drawbacks that can be used to store a session.

2.3 Data Sniffing

Most web browsers share access to a single browsing history and cache (file and DNS). This leads to history sniffing attacks, where a tracker can determine if the user has lately visited some other unconnected webpage.

- **Cache Timing**: The tracker can obtain this information calculating the time difference between the execution of certain operations related to data caching [18, 19]. This is possible because all web browsers implement many types of caching and the time needed to obtain the data are related to the browsing history of the user.
- **Information Leakage**: Sometimes an attacker does not need to perform any type of privacy violations by himself, because some trackers send their information via HTTP (e.g., using 1 pixel x 1 pixel transparent graphic images). Sniffing the traffic on the network would allow to obtain all the data that is being sent. Another leakage attack, exploits the fact that browsers display links in a different way if the webpage was previously accessed [20]. Using JavaScript, the tracker just needs to create a hidden link to the target webpage and then, making use of the browser's DOM, check how the link was presented. Depending on the result, it is possible to determine if the website is in the user's history.

2.4 Discussion

All the techniques described are somehow used in the wild, but their biggest limitations resides on the fact that each of them gives the tracker only some specific information about the user. Independently, these techniques are not as dangerous as they can be if combined, because trackers only get a partial overview. Data exchange between trackers that use different techniques is indeed, one of the main problems.

In the fast changing environment of web development, these attacks may not work if some specifications or properties of HTML5 and Javascript vary. Moreover, modifications related to regulations in data collections and privacy could invalidate their use.

3 Implementations in the wild

There are two different groups that implement privacy attacks in order to achieve specific objectives. Some of them use the extracted information to improve the quality of their service, others use the data with possible malicious intentions. Regardless of the objective, this information is being obtained without explicit acceptance of the end-user.

3.1 Advertising and Analytics Services

These services provide tools for websites to figure out the preferences of visitors, indicating demographics, browser, operating system, views and interactions [21]. They create usage profiles of the websites a user interacts with over time.

Although these implementations can differ from service to service, nearly all have adopted one of the two typical models. Some offer analytics as a paid service; they cannot use any client's analytics information and they protect the obtained information. Others offer a free analytics service, but they use the obtained data for ad targeting, market understanding or any other purpose. Advertising companies do not always depend on the data sold by the analytics services. They use their own techniques to understand the user. Information transference between banners is one of the most used techniques [22].

3.2 Self Implemented

We tend to think that websites only use pre-packaged solutions like the ones previously commented to obtain information of their users. However, some of them construct their own implementations [23]. Sometimes they are even obfuscated to evade detection systems. As these methods do not follow any specific information flow, they are much more difficult to detect and stop.

4 Standardizations

One possible solution to the problem of online privacy attacks is making a standard to control the information that is being transmitted. Two main projects have been advanced for giving users control over their personal data: Do Not Track (DNT) [24] and Platform for Privacy Preferences (P3P) [25].

4.1 Do Not Track

It is a proposal that combines technology and policies in order to send user's preferences on web tracking. This information is sent in a HTTP header, DNT. All modern browsers (Chrome, Firefox, Opera, Safari and Internet Explorer) support a Do Not Track opt-out preference (i.e., DNT: 1 header). The policy also indicates that websites must stop tracking the user for whatever reason when they receive a Do Not Track header.

4.2 Platform for Privacy Preferences

This project facilitates websites the task of communicating their privacy habits in a standard format that can be automatically obtained and understood by user agents. Users have the possibility of coming to a decision based on the privacy practices indicated by the website [26]. Thanks to that, users do not need to read the privacy policies of all the webpages they access, they just need to read it's practices. Sites implementing these policies have to make their habits public. Browsers can help the user to interpret those privacy habits with user-friendly interfaces.

4.3 Discussion

Although many stakeholders (policy makers, consumer advocates and researchers) think that Do Not Track could decidedly reduce tracking and data collection on the web, as the final decision of taking it into account only resides in websites, it is not followed as expected [27].

The case of P3P is similar, due to the lack of support from current browsers for the implementation, the P3P Specification Working Group suspended the project.

5 Defense methods

Accessing certain websites can lead to information leakage that could harm the user on many levels [28], including their own privacy. Although some people think that they may not be the final objective of any privacy attack, information of millions of users is being collected everyday. In order to prevent some of those leaks, we describe the two main approaches proposed in the literature.

5.1 Identification and Control

To protect the end-user from the different privacy attacks, there are fully functional anti-tracking web browsers that implement a precise and general information analysis and control [29, 30]. In order to make browsing as normal as possible, they try to have a low performance overhead. There is some research that focuses all the analysis in user's browsing [31]. In that way, all the accessed

pages could be analyzed without exception, taking into account that the user is the weakest link in the security chain. Applying taint analysis or dynamic controls and making use of determine policies, it is possible to detect privacy violations [32, 33].

5.2 Spoofing and Configurations

Spoofing the browser profile can guard against many attempts of user tracking. Although the best option is that all users have the same browser profile, it is impossible. Spoofing the data and having random browser profiles could help, because it eradicates the possibility of identifying a user for the uniqueness of his browser. Some of these properties are browser, platform, time zone or screen resolution.

Regarding to possible configurations that could avoid some privacy attacks, the most effective but less appropriate one is disabling Javascript. Most of the attacks described use or depend somehow in Javascript. Another option to protect the web privacy, is to browse in temporary modes such as private or guest mode, so the browser does not save or cache what you visit and download [34]. Some other key points are disabling certain font sets, cookies and plugins by default or blocking the requests to websites listed as tracking servers.

5.3 Discussion

The main problem of identification and control methods is that they only take into account certain fields and privacy attacks, forgetting about the rest. Understanding and controlling every type of privacy attack would enormously improve web-browsers. Nevertheless, the biggest disadvantage of a general control, as taint analysis, is that they are not computationally efficient.

Regarding configurations, disabling Javascript would prevent many tracking approaches, but it would stop many websites from rendering correctly. Disabling other secondary aspects used in privacy attacks can be a better choice because the number of websites that rely on them is much smaller. Finally, spoofing could be counterproductive because these attempts to hide the identity may be fingerprintable [35].

6 Regulations

After understanding the magnitude of the problem, we should know the existing regulations of the area in the United States and European Union [36]. It was not until recently that these regulations appeared in order to restrict the ability of large-scale collection of personal data [37].

6.1 United States

One of the missions of the Federal Trade Commission (FTC) is the promotion of consumer protection. They can only prevent practices of businesses that are

either unfair or deceptive under 15 U.S.C. § 45. First violations will incur on a small payment, but subsequent violations get big monetary penalties.

On 2012 the FTC issued its final report [38] establishing four best practices for companies to protect the privacy of all American consumers and give them the possibility to have more control of tracking options and personal information collection. The report expands on a preliminary report released in 2010 [39], which proposed a framework for consumer privacy control because of the new technologies that allow for information collection that is often not perceivable by consumers. The objective is to balance the personal data of consumers with innovation.

6.2 Europe

The Directive 2002/58/EC on Privacy and Electronic Communications, also known as E-Privacy Directive, indicates that the use of electronic communications networks to store information or to gain access to information stored in the terminal equipment of a user is only allowed on condition that the user concerned is provided with clear and comprehensive information about the purposes of the processing, and is offered the right to refuse such processing [40]. If the above indicatations are not met, penalties could be up to 2% of the revenue.

The Article 29 Working Party (WP29) addresses the topic of device fingerprinting in the Opinion 9/2014 , which extends over the previous Opinion on Cookie Consent Exemption [41], and indicates that websites cannot process device fingerprints which are generated through the gaining of access to or the storing of information on the users terminal device if there is not a explicit consent of the user (unless some specific exemptions) [42].

6.3 Self-regulation

In 2009, many of the largest advertising and marketing companies and associations, supported by the Council of Better Business Bureaus, created a self-regulatory program with the principal objective of giving total control over the collection and use of private data to the users [43]. Websites should have clear options regarding to the data collection and use, letting the user decide if they want that collection or not. There should also be a limit on the specific data type obtained if it is sensitive information. Until that moment, all the different actors worked interdependently in this area. Nevertheless, it is only indicated for the data collection used to predict user interests to deliver online advertising. These principles do not apply to websites that collect that information for its own uses.

2 years later, the Digital Advertising Alliance (DAA) announced and expansion of the program in order to include the non-advertising businesses to the self-regulation [44]. These new principles prohibit third parties to collect, use or transfer any multi-site information. However, these data was mostly covered in the areas of insurance, credit, employment or health.

6.4 Discussion

Although many regulations exist, there is not a continuous control of the websites to check if they are actually following them. Creating a organization responsible for this would secure the compliance of regulations and therefore improve the privacy control of the users.

7 Conclusion

This paper analyzed and discussed the different factors related to online web tracking as of early 2015. Privacy is an area in continuous evolution that directly interferes in the end-users and need to be addressed in order to protect them.

We hope that the survey presented here provides security and privacy researchers with a good background in order to contribute to the field. Future work is oriented to developing new detection methods for privacy violations in order to enhance the results and the system performance of existing ones.

Acknowledgment. This research was partially supported by the Basque Government under the pre-doctoral grants given to Iskander Sánchez-Rola and Xabier Ugarte-Pedrero.

References

1. Stevenson, A.: Oxford Dictionary of English. Oxford Dictionary of English. OUP Oxford (2010)
2. Milanovic, M.: Human rights treaties and foreign surveillance: Privacy in the digital age. Harvard International Law Journal, Forthcoming (2014)
3. Bernal, P.: Internet Privacy Rights: Rights to Protect Autonomy. Volume 24. Cambridge University Press (2014)
4. Squicciarini, A.C., Paci, F., Sundareswaran, S.: Prima: a comprehensive approach to privacy protection in social network sites. annals of telecommunications-annales des télécommunications **69**(1-2) (2014) 21–36
5. Wang, Y., Nepali, R.K., Nikolai, J.: Social network privacy measurement and simulation. In: Computing, Networking and Communications (ICNC), 2014 International Conference on, IEEE (2014) 802–806
6. Cecere, G., Rochelandet, F.: Privacy intrusiveness and web audiences: Empirical evidence. Telecommunications Policy **37**(10) (2013) 1004–1014
7. Hayes, C.M., Kesan, J.P., Bashir, M., Hoff, K., Jeon, G.: Informed consent and privacy online: A survey. Available at SSRN 2418830 (2014)
8. Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F., Preneel, B.: Fpdetective: Dusting the web for fingerprinters. In: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, ACM (2013) 1129–1140
9. Eckersley, P.: How unique is your web browser? In: Privacy Enhancing Technologies, Springer (2010) 1–18

10. Fifield, D., Egelman, S.: Fingerprinting web users through font metrics. In: Proceedings of the 19th international conference on Financial Cryptography and Data Security. (2015)
11. Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., Diaz, C.: The web never forgets: Persistent tracking mechanisms in the wild. In: Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS 2014). (2014)
12. Mowery, K., Shacham, H.: Pixel perfect: Fingerprinting canvas in html5. Proceedings of W2SP (2012)
13. Ayenson, M., Wambach, D.J., Soltani, A., Good, N., Hoofnagle, C.J.: Flash cookies and privacy ii: Now with html5 and etag respawning. Social Science Research Network (2011)
14. Atterer, R., Wnuk, M., Schmidt, A.: Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In: Proceedings of the 15th international conference on World Wide Web, ACM (2006) 203–212
15. Keromytis, A.: Darpa, active authentication program. `http://www.darpa.mil/our_work/i2o/programs/active_authentication.aspx`
16. Soltani, A., Canty, S., Mayo, Q., Thomas, L., Hoofnagle, C.J.: Flash cookies and privacy. In: AAAI Spring Symposium: Intelligent Information Privacy Management. (2010)
17. West, W., Pulimood, S.M.: Analysis of privacy and security in html5 web storage. Journal of Computing Sciences in Colleges **27**(3) (2012) 80–87
18. Felten, E.W., Schneider, M.A.: Timing attacks on web privacy. In: Proceedings of the 7th ACM conference on Computer and communications security, ACM (2000) 25–32
19. Focardi, R., Gorrieri, R., Lanotte, R., Maggiolo-Schettini, A., Martinelli, F., Tini, S., Tronci, E.: Formal models of timing attacks on web privacy. Electronic Notes in Theoretical Computer Science **62** (2002) 229–243
20. Weinberg, Z., Chen, E.Y., Jayaraman, P.R., Jackson, C.: I still know what you visited last summer: Leaking browsing history via user interaction and side channel attacks. In: Security and Privacy (SP), 2011 IEEE Symposium on, IEEE (2011) 147–161
21. Altaweel, I., Cabrera, J., Choi, H.S., Ho, K., Good, N., Hoofnagle, C.: Web privacy census: Html5 storage takes the spotlight as flash returns
22. Roesner, F., Kohno, T., Wetherall, D.: Detecting and defending against third-party tracking on the web. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. NSDI'12, Berkeley, CA, USA, USENIX Association (2012) 12–12
23. Jang, D., Jhala, R., Lerner, S., Shacham, H.: An empirical study of privacy-violating information flows in javascript web applications. In: Proceedings of the 17th ACM conference on Computer and communications security, ACM (2010) 270–283
24. Narayanan, A., Mayer, J.: Do not track, universal web tracking opt out. `http://donottrack.us`
25. World Wide Web Consortium: Platform for privacy preferences (p3p) project. `http://www.w3.org/P3P`
26. Byers, S., Cranor, L.F., Kormann, D., McDaniel, P.: Searching for privacy: Design and implementation of a p3p-enabled search engine. In: Privacy Enhancing Technologies, Springer (2005) 314–328
27. Mayer, J.: Tracking the trackers: early results. `http://cyberlaw.stanford.edu/blog/2011/07/tracking-trackers-early-results` (2011)

28. Teltzrow, M., Kobsa, A.: Impacts of user privacy preferences on personalized systems. In: Designing personalized user experiences in eCommerce. Springer (2004) 315–332
29. De Groef, W., Devriese, D., Nikiforakis, N., Piessens, F.: Flowfox: a web browser with flexible and precise information flow control. In: Proceedings of the 2012 ACM conference on Computer and communications security, ACM (2012) 748–759
30. Pan, X., Cao, Y., Chen, Y.: I do not know what you visited last summer: Protecting users from third-party web tracking with trackingfree browser. In: NDSS: Proceedings of the Network and Distributed System Security Symposium. (2015)
31. Hedin, D., Birgisson, A., Bello, L., Sabelfeld, A.: Jsflow: Tracking information flow in javascript and its apis. In: Proc. 29th ACM Symposium on Applied Computing. (2014)
32. Sen, K., Kalasapur, S., Brutch, T., Gibbs, S.: Jalangi: A selective record-replay and dynamic analysis framework for javascript. In: Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, ACM (2013) 488–498
33. Chugh, R., Meister, J.A., Jhala, R., Lerner, S.: Staged information flow for javascript. In: ACM Sigplan Notices. Volume 44., ACM (2009) 50–62
34. Aggarwal, G., Bursztein, E., Jackson, C., Boneh, D.: An analysis of private browsing modes in modern browsers. In: USENIX Security Symposium. (2010) 79–94
35. Nikiforakis, N., Kapravelos, A., Joosen, W., Kruegel, C., Piessens, F., Vigna, G.: Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In: Security and privacy (SP), 2013 IEEE symposium on, IEEE (2013) 541–555
36. Mayer, J.R., Mitchell, J.C.: Third-party web tracking: Policy and technology. In: Security and Privacy (SP), 2012 IEEE Symposium on, IEEE (2012) 413–427
37. Goldfarb, A., Tucker, C.E.: Privacy regulation and online advertising. Management Science **57**(1) (2011) 57–71
38. Federal Trade Commission: Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers". `https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers` (2012)
39. Federal Trade Commission: Protecting consumer privacy in an era of rapid change, a proposed framework for businesses and policymakers". `https://www.ftc.gov/reports/preliminary-ftc-staff-report-protecting-consumer-privacy-era-rapid-change-proposed-framework` (2010)
40. European Parliament: Directive 2002/58/ec. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:en:HTML` (2002)
41. Article 29 Data Protection Working Party: Opinion 04/2012 on cookie consent exemption. `http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2012/wp194_en.pdf` (2012)
42. Article 29 Data Protection Working Party: Opinion 9/2014 on the application of directive 2002/58/ec to device fingerprinting. `http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp224_en.pdf` (2014)
43. Digital Advertising Alliance: Self-regulatory principles for online behavioral advertising, behavioral advertising. `http://www.aboutads.info/resource/download/seven-principles-07-01-09.pdf` (2009)
44. Digital Advertising Alliance: Self-regulatory principles for multi-site data. `http://www.aboutads.info/resource/download/Multi-Site-Data-Principles.pdf` (2011)